The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0737-8831.htm

COLUMN: ARCHIVING IN THE NETWORKED WORLD

Archiving in the networked world: betting on the future

Michael Seadle
Humboldt-Universität zu Berlin, Germany

Received 12 March 2009
Revised 14 March 2009
Accepted 15 March 2009

Abstract

Purpose – The goal of this column is not to argue the pros and cons of digital archiving, or to propose solutions to its problems, but to describe it as a research subject and a social phenomenon.

Design/methodology/approach – This column relies on cultural anthropology, in particular the approach that Clifford Geertz championed, and for cultural anthropology, language and its social context matter.

Findings – Archiving systems abound with competing claims about effectiveness. Transparency and evidence of public testing is rare, with a few exceptions. The lack of public testing does not mean that systems do less than they claim, but it does mean that libraries, archives and museums need to press for proof if they want to have confidence in the product.

Originality/value – When betting on the future, these cannot be certainty, but bets placed should be based on knowledge.

Keywords Digital libraries, Library and information networks, Museums

Paper type General review

Introduction

This new *Library Hi Tech* column is about long-term digital archiving, a topic that seems increasingly urgent as more and more information appears in digital form. The goal of this column is not to argue the pros and cons of digital archiving, or to propose solutions to its problems, but to describe it as a research subject and a social phenomenon.

I am not neutral on the subject. It quickly will become obvious to readers that I believe that digital archiving is a better long-term solution for preserving information than paper or microfilm or other contemporary methods. I also believe it has flaws. An important leitmotiv is that research on long-term digital archiving is in its infancy.

Fragility

The collapse of the archive in Cologne in March 2009 is merely one more example of the fragility of physical documents. The building contained a broad range of unique materials. The news magazine *Der Spiegel* lists the following:

> *Mehr als 65,000 Urkunden aus dem Raum Köln ab dem Jahr 922, 104,000 Karten und Pläne, 50,000 Plakate und rund eine halbe Million Fotos. Zudem sind dort 780 Nachlässe und Sammlungen, unter anderem von Literaturnobelpreisträger Heinrich Böll gelagert worden.*

Library Hi Tech
Vol. 27 No. 2, 2009
pp. 319-325

0737-8831
DOI 10.1108/07378830910968326

> More than 65,000 documents from the Cologne area since the year 922, 104,000 maps and plans, 50,000 posters and about a half million photos. In addition, it stored 780 donations and collection from, among others, the Nobel Prize winning novelist Heinrich Böll (my translation – Spiegel Online, 2009).

Some documents may be recovered from under the rubble, just as some works were rescued from the fire at the Anna Amalia Library in Weimar in 2005. A great many will be lost. Those who deposited works at the archive made a bet that that physical location was a safe place to store important works for the future. They lost the bet.

Method

Since this is the first article in this column series, it is important to describe the scholarly method that lies behind it, even though I have described my general approach in detail elsewhere (see Seadle, 2000).

This column relies on cultural anthropology, in particular the approach that Clifford Geertz championed, and for cultural anthropology, language and its social context matter. While I have not lived among the natives of Indonesia or Morocco, I have lived and worked regularly among librarians and archivists since 1976. I have learned how they think and I have learned the languages they speak. I have even gone through the ritual initiation ceremony of acquiring a Masters degree accredited by the American Library Association. I have also worked with and observed museum staff, particularly those engaged with historical objects, though I am less acculturated to that world. For much of the time since 1976 I have also worked as and with computer professionals. While I am no computer scientist in the strict academic sense of having completed a degree in the field, I have worked long enough and intensely enough with assembly language, operating system architecture, internet protocols and system design to hold rational discussions on computing topics.

When dealing with archiving in the networked world, the speakers of these two language groups – those of librarians, archivists, and museum staff on the one hand, and those of various types of computer professionals on the other – need to interact and to understand one another. Just as traditional anthropology translates one human culture to another, so do I attempt in this column to translate between these professional groups. I used the phrase "language groups" deliberately. The language of archivists is by no means the same as that of librarians or museum staff, and the range of speech and concept is, for those who code Java interfaces, markedly different from those who optimize input/output systems. They function much as other living language groups do, with structural similarities, common words, and many false cognates.

My education in digital archiving comes in part from long years of computer center experience, and in part from a close multi-year relationship with LOCKSS (Lots of Copies Keep Stuff Safe) from Stanford University. Knowing one archiving system offers a potential for bias, but bias is inevitable in cultural anthropology, since we cannot escape coming from a particular culture. Part of anthropological training is to recognize influences from our own background and to adjust for them. Ignorance of archiving systems would be no improvement.

Throughout this column, as in most of my other scholarly writing, I use the first person singular. Anthropology is an empirical field and it is important to distinguish which data come from my observations and which from other sources.

Copies

The word "copy" evokes different feelings among librarians, archivists and computer professionals. One of the first questions that appeared on the German e-mail list for librarians was whether works in the collapsed archive had been microfilmed (InetBib, 2009). At the time of writing this column, it is not clear what the answer is, but the question is important. Librarians work constantly with physical copies of published works and recognize that a broad distribution of copies is important for a work's survival. Librarians have little hesitation about discarding older editions of books when a new one appears, unless the new version leaves out a famous preface or famous graphics. For librarians, content matters more than the exact form on the page.

For archivists the actual physical original has a value and meaning that no copy can replace. Archivists deal normally with unique materials, not works published in hundreds or thousands of copies. Some archivists do make copies to protect their original documents from excessive use. They are aware, as are librarians, that use damages works and that for heavily used works the ordinary reader represents a greater hazard than fire, flood, or building collapse. Even well-intentioned users may leave oil from their hands on a page unless they are forced to wear white gloves, or may damage the spine of a bound work merely by opening it. Archivists do not view copies as substitutes, and they care as deeply as art museum curators about the provenance and chain-of-ownership of works in their possession. Proof of authenticity matters for archives, while for librarians the authenticity of a published book is accepted on the basis of the name on its cover.

Computer professionals encourage copies. They want backups of every file and generally feel safest with multiple copies in multiple locations. They work in an environment where physical media have a short lifespan and low reliability, but where a copy is literally indistinguishable from an original. Digital works cannot in fact be used without copies. To read a work from a disk it must first be copied into working memory. Reading a work from a networked location means making a temporary copy on the local (client) machine. Use also actually assists in preserving a digital work, because use can signal a media problem that requires replacement from a backup. Mere usage does no damage to a server copy. Reading the same spot again and again on a disk does not cause any wear in that location because the read-head never actually touches the spot – if it does, that means a disk crash and the whole disk is ruined. The question of which copy represents the authentic original is, for a computer file, essentially meaningless. No computer professional worries about whether the Cologne archive lost "original" digital files, only whether the files were backed up and kept in an off-site location.

Microfilm

In the 1970s and 1980s librarians tended to regard microfilm as a reasonable medium for addressing the problem of fragile works that were no longer available in the open market. After librarians became aware of the acid paper problem, a number of libraries attempted systematic microfilming projects that swept through the stacks filming large numbers of works. The film originals were then sent to climate-controlled vaults, except when new copies were needed to make prints. I sent fragile works out for filming in the 1980s and had the sense that I was employing a cutting-edge technology. When the Center for Research Libraries ran out of space for newspapers in the late

1970s, I talked with its then president about filming the papers. It would preserve them and make them much easier to send copies to member libraries that wanted to borrow them. The flaw in my idea was, the president explained, that no donor would pay to have his name on a box of microfilm, but some rich people would pay to have their name on a building.

I became aware of other problems over time. It was true that black and white silver halide film should (in theory) last hundreds of years, but achieving the ideal lifespan required particular environmental conditions. At a time when nuclear war seemed imminent, storage in remote, climate controlled mountain caves offered protection against both bombs and ordinary deterioration. Such storage had a cost, of course. Microfilm readers were another cost. The market for them was small. Few scholars ever purchased private microfilm readers in any quantity. A few companies did, but not enough to drop the cost to commodity levels. Cost was, however, less of a problem than the users themselves, who generally hated the machines as both unreliable and hard on the eyes. They also discovered that pages might be filmed askew or left out accidentally. In those days images became visible only after the film was developed and could not easily be remedied short of creating a new master. Every new master also meant some small loss of quality, too little to worry about in any one lifetime, but measurable over millennia.

The Cornell/Xerox/Commission on Preservation and Access digitization project in 1990 aimed to create digital images that would meet digitization standards for preservation. (Kenney and Personius, 1991) Since the image density standards of the time were microfilm-based, they were hard for contemporary scanners to achieve without interpolation. In fact, merely capturing the information on a page of print required no high standard of density, but capturing the exact page with all of its flaws, including discoloration due to acid paper deterioration, required more. Librarians began to ask themselves what the goal was: information capture or page reproduction that showed the original object as exactly as possible? (Kenney and Chapman, 1996). Clearly, if a famous mathematician made marginal notes on a page, the notes mattered, but generally damage was just damage and not an intellectual contribution for eternity.

Color presented a problem for microfilm. Color film was more expensive and, until the 1980s, it tended to fade rapidly over time. Trueness remained a problem in color too: different film producers and different development techniques affected the color of the end result. Even today, the chemical composition of color film lacks the long-term stability of silver-halide. Meanwhile, color publication has become increasingly common thanks to inexpensive digital printing techniques. Some experimentation was done with computer-output-microfilm as a way of combining the relative accuracy of a digital image with the theoretical durability of microfilm, but it never caught on (Kenny, 1997).

Why did microfilm fail to dominate the preservation market once digital scanning became broadly available? Microfilm had a significant head start in the market, which usually gives a significant advantage, but microfilm's share of the market for preservation copies has fallen steadily while digitization has grown. There are a variety of reasons, perhaps chief among them the unpopularity of both microfilm and microfiche among users. Relatively few outside the library community care about having a physical surrogate that may possibly survive a nuclear holocaust in some mountain bunker, but is fundamentally useless until converted to some more

user-friendly form. And if the mountain bunker floods or its roof falls in, that physical copy is lost forever. Such events are perhaps unlikely, but so was the collapse in Cologne.

Probabilities

Probability is a word that discomforts many archivists when speaking about long-term archiving. The discourse in that field revolves around the concept of secure or trustworthy locations. When dealing with a single unique work, the goal of a secure location is to maximize the probability that the location will be able to keep the work safe and undamaged. Those probabilities are hard to judge for a single physical location, though not impossible. Statistics exist in most western countries about how often buildings of particular types suffer from fire, flood, or physical collapse. These events happen even to locations believed to be secure: the Anna Amalia Library fire, the Colorado State Library flood, and the Cologne archive collapse.

Probabilities affect physical objects in other ways too. Any time users have access to a work, it faces the danger of damage. It is no secret that users have razored out pages from works in open stacks or stolen them whole. JSTOR famously had to search for copies of some articles from multiple libraries to find an intact version for digitization. The probability of a user damaging an original Luther Bible is distinctly smaller than the probably of damage to a work in an open stacks, because no one would leave the Luther Bible in the hands of a user unsupervised. Librarians and archivists *de facto* grade the acceptable risk for works by how they expose the works to use. Usage risk comes in addition to the risk of building damage. In other words, the two risks should be added together.

Computer professionals view risk differently and speak about it more explicitly, partly because risk in the digital environment is more obvious. No one imagines that a hard disk will last 200 years, which was, until recently, the building standard for construction in Germany. A disk lifespan of year or two is more common. This means that computing professionals are more accustomed to addressing the probability of risk and of compensating appropriately. Even so, computing centers do lose data occasionally. In my years as a computing professional I have dealt with unreadable backup tapes and formats where I had to do reverse-engineering to recover information. A rule-of-thumb in computer centers is multiple copies on multiple media types for really valuable data.

A standard criticism of digital copies is that the risk of the failure of any one medium or any one format is higher than the risk for any single paper-based or microfilm-based work. This is certainly true, but from a computing perspective, irrelevant as long as appropriate back-up procedures are followed. The key question from a computing perspective is how to minimize the risk with multiple copies and routine integrity testing, and this is the fundamental principle behind the LOCKSS system. The probabilities can be calculated and tested in a digital environment and it seems clear that more copies with more checking enhances security.

Long-term archiving, whether physical or digital, is a bet on the future, and making a rational bet involves a probability analysis whose validity depends on the kind and quality of data fed into it. This is one of the important research areas for archiving that has received relatively little attention.

Digital preservation

Digital preservation strategies vary by object-types. For discussion purposes these types might be organized into the following categories of digital copies:

(1) static two-dimensional works (e.g. print publications, whether paper-based or electronic);

(2) static three-dimensional works (e.g. museum objects);

(3) multimedia (e.g. video or film);

(4) interactive works (e.g. computer games); and

(5) raw data (e.g. digital feeds from a scientific experiment).

Four of these five types are today born digital, including essentially all print publications, which invariably go through a digital editing phase even if a first draft was created with pen and ink. Museum objects alone are an exception. Despite this variety of types, the discourse about digital preservation in the library and archive literature still focuses strongly on paper-based works because those are the media with which librarians and archivists traditionally identify. If libraries continue to limit their interest, computer centers or other organizations may step in and make libraries irrelevant for preserving some of the fastest growing areas for communicating information.

I know a number of librarians who say they would welcome this. Their point is that preserving two-dimensional works is hard enough, but not everyone agrees. Herbert van de Sompel remarked at the 2009 Bielefeld Conference that European libraries are "more focused on established literature" than American libraries. He also strongly recommended taking an interest in the data behind published research results, which could in many cases go into institutional repositories. These data are not generally part of publisher copyright agreements today, which removes one significant barrier to action.

Copyright ownership affects all types of digital archiving and deserves more extensive discussion. Since libraries generally license rather than own digital works, future access may depend on whether they negotiate a contract that includes archiving rights and access, if and when a publisher goes out of business. Many questions remain unexplored here. No one has tested the possibility that a dark archive purely for long-term preservation purposes would not violate copyright law, if the copies remained inaccessible until they became public domain. Digital rights management software might also play a role, as Dörte Böhner has suggested (Böhner, 2008).

Conclusion

The attention given to archiving in the networked world is increasing in part because the amount of information available on the internet has reached a level where even ardent advocates of paper publication cannot reasonably deny its importance to the scholarly world. Archiving systems abound with competing claims about effectiveness. Transparency and evidence of public testing is rare, with a few exceptions. The lack of public testing does not mean that systems do less than they claim, but it does mean that libraries, archives and museums need to press for proof if they want to have confidence in the product.

Research topics abound in the area of long term digital archiving. A few have been mentioned already and more need to be explored. This column will look at new research as information about it becomes available. One area where my colleague Elke Greifeneder and I are conducting is on reading, specifically on whether people read differently when interacting with different visual and physical formats. Today, digital archiving often preserves the page view with a PDF. I long thought this important, but such formats could become as archaic as scrolls in the age of codexes, if eBook readers become common.

When betting on the future, we cannot have certainty, but we should place our bets based on knowledge.

References

Böhner, D. (2008), "Digital rights description as part of digital rights management: a challenge for libraries", *Library Hi Tech*, Vol. 26 No. 4, pp. 598-605, available at: www.emeraldinsight.com/Insight/viewContentItem.do;jsessionid = 56A90F22905CB1D492501D4E6E138AB8? contentType = Article&contentId = 1753933 (accessed January 7, 2009).

InetBib (2009), *[InetBib] R: Koelner Katastrophe*, available at: www.ub.unidortmund.de/listen/inetbib/date1.html (accessed March 9, 2009).

Kenney, A. (1997), "The Cornell Digital to Microfilm Conversion Project: final report to NEH", *RLG DigiNew*, Vol. 1 No. 2, available at: www.rlg.org/preserv/diginews/diginews2.html

Kenney, A. and Chapman, S. (1996), *Bitonal Scanning Means for Benchmarking Resolution Requirements in Digital Imaging for Libraries and Archives*, Cornell University Libraries, Ithaca, NY, pp. 7-9.

Kenney, A. and Personius, L. (1991), *The Cornell/Xerox/Commission on Preservation and Access Joint Study in Digital Preservation*, Washington, DC: Commission on Preservation and Access, available at: www-cpa.stanford.edu/cpa/reports/joint/

Seadle, M. (2000), "Project ethnography: an anthropological approach to assessing digital library services", *Library Trends*, Vol. 49 No. 2.

Spiegel Online (2009), "Einsturz des Kölner Stadtarchivs: 'Alle weg, alle raus!'", available at: www.spiegel.de/panorama/0,1518,611169,00.html (accessed March 4, 2009).

Corresponding author
Michael Seadle can be contacted at: seadle@ibi.hu-berlin.de